Bayesian Discrimination

by

Seymour Geisser
University of Minnesota

Technical Report No. 354

July, 1979

Bayesian Discrimination

by

Seymour Geisser
University of Minnesota

1. Introduction

The complementary problems of allocation and separation of structured populations are reviewed and amplified. In either case, we assume we have two or more identifiable populations whose distribution functions for a set of manifest variables are known up to some specifiable parameters. An identifiable sample is drawn from the populations. In one case future observations to be generated, or observations possibly already in hand but whose latent identity is unknown, require labeling, diagnosis or allocation. In the second case, we require some simple functions (discriminants) which maximally distinguish or separate these populations. This is attempted in order to throw some light on relevant issues or to formulate hypotheses concerning these populations. Sometimes the goal is to make high dimensional data more immediately accessible and more manageable by severely reducing their dimensionality yet retaining a large degree of the total information available in the data.

We first describe a general Bayesian procedure for allocation and then give applications for the most popular of models in this area, the multivariate normal one. The problem of separation from a Bayesian viewpoint is then presented.

Often both allocation and separation are part of the same study and some compromise solutions, which can serve in a near optimal manner for both purposes, are obtained and applied to multivariate normal populations. A sample reuse procedure in conjunction with a semi-Bayes approach which

- - - - - - - - - - -

*Work was supported in part by NIH-GM-25271.

is useful for selecting the appropriate allocatory/separatory model is also presented. Further areas for examination via the Bayesian approach are proposed.

2. Bayesian Allocation

Suppose we have k populations π_i, $i = 1,\ldots,k$, each specified by a density $f(\cdot|\theta_i,\psi_i)$, where θ_i is the set of distinct unknown parameters of π_i; ψ_i is the set of distinct known parameters of π_i; X_i is the data obtained on π_i based on N_i independent (vector) observations; and z is a new (vector) observation to be assigned which has prior probability q_i of belonging to π_i, $\Sigma_{i=1}^{k} q_i = 1$.

Further, let $\theta = \cup_{i=1}^{k} \theta_i$, $\psi = \cup_{i=1}^{k} \psi_i$, i.e., the total set of distinct unknown and known parameters, respectively, and $g(\theta|\psi)$ be the joint prior density of θ for known ψ. Let $L(X_i|\theta_i,\psi_i)$ be the likelihood of the sample obtained from π_i with the joint likelihood obtained on $\pi_1,\ldots,\pi_k$ given by

$$L(X|\theta,\psi) = \prod_{i=1}^{k} L(X_i|\theta_i,\psi_i), \tag{2.1}$$

where X represents the set of all the data samples $X_1,\ldots,X_k$ often referred to as the training sample. Hence the posterior density, when it exists, is

$$p(\theta|X,\psi) \propto L(X|\theta,\psi)g(\theta|\psi), \tag{2.2}$$

from which we may obtain the predictive density of z on the hypothesis that it was obtained from π_i, which results in

$$f(z|X,\psi,\pi_i) = \int f(z|\theta_i,\psi_i,\pi_i)p(\theta|X,\psi)d\theta\ . \tag{2.3}$$

Occasionally it is more convenient to express the above equation in the following manner:

$$f(z|X,\psi,\pi_i) = \int f(z|\theta_i,\psi_i,\pi_i)p(\theta_i|X,\psi)d\theta_i, \tag{2.4}$$

where

$$p(\theta_i|X,\psi) = \int p(\theta|X,\psi)d\theta_i^c \tag{2.5}$$

and θ_i^c is the complement of θ_i, $\theta_i^c \cup \theta_i = \theta$. We then calculate the posterior probability that z belongs to π_i,

$$\Pr\{z \in \pi_i | X,\psi,q\} \propto q_i f(z|X,\psi, \pi_i), \tag{2.6}$$

where q stands for $(q_1,\ldots,q_k)$. For allocation purposes we may choose to assign z to that π_i for which (2.6) is a maximum, if we ignore the differential costs of misclassification. We could also divide up the observation space of z into sets of regions $R_1,\ldots,R_k$, where R_i is the set of regions for which $u_i(z) = q_i f(z|X,\psi, \pi_i)$ is maximum and use these as allocating regions for future observations. We may also compute "classification errors," based on the predictive distributions, which are in a sense a measure of the discriminatory power of the variables or characteristics. If we let $\Pr\{\pi_j|\pi_i\}$ represent the predictive probability that z has been classified as belonging to π_j when in fact it belongs to π_i, then we obtain

$$\Pr\{\pi_i|\pi_i\} = \int_{R_i} f(z|X,\psi,\pi_i)dz, \tag{2.7}$$

$$\Pr\{\pi_j|\pi_i\} = \int_{R_j} f(z|X,\psi,\pi_i)dz \quad (i \neq j), \tag{2.8}$$

$$\Pr\{\pi_i^c|\pi_i\} = \left(1 - \int_{R_i} f(z|X,\psi,\pi_i)dz\right), \tag{2.9}$$

where π_i^c stands for all the populations with the exception of π_i. Then the predictive probability of a misclassification is

$$\sum_{i=1}^{k} q_i \Pr\{\pi_i^c|\pi_i\} = 1 - \sum_{i=1}^{k} q_i \Pr\{\pi_i|\pi_i\}. \tag{2.10}$$

Prior to observing z, the smaller the predictive probability of a misclassification the more confidence we have in the discriminatory variables. However, once z has been observed and if our interest is only

in the particular observed z, the misclassification errors are relatively unimportant, but what is important is (2.6), i.e., the posterior probability that Z belongs to π_i. Nevertheless, before any observations are inspected for assignment, the error of classification can be of value in determining whether the addition of new variables or the deletion of old ones is warranted.

In many situations the q_i's are also unknown. First we consider that the sampling situation was such that we have the multinomial density for the N_i's (where throughout what follows $N_k = N - N_1 - \ldots - N_{k-1}$, and $q_k = 1 - q_1 - \ldots - q_{k-1}$). Thus the likelihood for the observed frequencies in the training sample is

$$L(q_1,\ldots,q_{k-1}) \propto \prod_{j=1}^{k} q_j^{N_j}. \tag{2.11}$$

If we assume that the prior probability density of the q_i's is of the Dirichlet form

$$g(q_1,\ldots,q_{k-1}) \propto \prod_{j=1}^{k} q_j^{\alpha_j}, \tag{2.12}$$

we obtain the posterior density of the q_i's,

$$p(q_1,\ldots,q_{k-1}\ N_1,\ldots,N_{k-1}) \propto \prod_{j=1}^{k} q_j^{N_j+\alpha_j}. \tag{2.13}$$

Further

$$p(q_1,..,q_{k-1}|z,N_1,..,N_{k-1}) \propto p(q_1,..,q_{k-1}|N_1,..,N_{k-1})f(z|q_1,..,q_{k-1},X,\psi) \tag{2.14}$$

where

$$f(z|q_1,\ldots,q_{k-1}) = \sum_{j=1}^{k} q_j f(z\ X,\psi,\pi_j), \tag{2.15}$$

whence we obtain the posterior probability no longer conditioned on q,

$$\Pr(z\in\pi_i|X,\psi) = \int..\int \Pr(z\in\pi_i|X,\psi,q)\,p(q_1,..,q_{k-1}|z,N_1,..,N_{k-1})\,dq_1..dq_{k-1}$$

$$= \frac{(N_i+\alpha_i+1)\,f(z|X,\psi,\pi_i)}{\Sigma_j (N_j+\alpha_j+1)\,f(z|X,\psi,\pi_j)} \qquad (2.16)$$

In the second situation we assume that the N_i's were chosen and not random variables. This is tantamount to assuming that $N_i = 0$ for all i as regards the posterior distribution of the q_i's, resulting in

$$\Pr(z\in\pi_i|X,\psi) \propto (\alpha_i+1)\,f(z|X,\psi,\pi_i) \qquad (2.17)$$

The α_i's may be regarded as reflecting previous frequencies or intuitive impressions about the frequencies of the various π_i's. If there is neither previous data nor any other kind of prior information the assumption $\alpha_i = \alpha$ for all i leads to the same result that we would obtain had we assumed that the k populations were all equally likely a priori, i.e. $q_i = 1/k$.

Suppose we wish to classify jointly n independent observations $z_1,\ldots,z_n$, each having prior probability q_i of belonging to π_i. We can then compute the joint predictive density on the hypothesis that $(z_1 \in \pi_{i_1},\ldots,z_n \in \pi_{i_n})$, where $i_1,\ldots,i_n$ are each some integer such that $1 \le i_j \le k$, $j = 1,\ldots,n$. Therefore,

$$f(z_1,\ldots,z_n|X,\psi,\pi_{i_1},\ldots,\pi_{i_n})$$

$$= \int p(\theta|\psi,X) \prod_{j=1}^{n} f(z_j|\theta_{i_j},\psi_{i_j},\pi_{i_j})\,d\theta$$

or

$$= \int p\left(\bigcup_{j=1}^{n} \theta_{i_j}\,|\psi,X\right) \prod_{j=1}^{n} f(z_j|\theta_{i_j},\psi_{i_j},\pi_{i_j})\,d \bigcup_{j=1}^{n} \theta_{i_j}\,, \qquad (2.18)$$

where

$$p\left(\bigcup_{j=1}^{n}\theta_{i_j}|\psi,X\right) = \int p(\theta|\psi,X)d\left[\bigcup_{j=1}^{n}\theta_{i_j}\right]^c . \tag{2.19}$$

This then yields the joint posterior probability

$$\Pr\{z_1 \in \pi_{i_1},\ldots,z_n \in \pi_{i_n}|X,\psi,q\} \propto \left(\prod_{j=1}^{n} q_{i_j}\right) f(z_1,\ldots,z_n|X,\psi,\pi_{i_1},\ldots,\pi_{i_n}). \tag{2.20}$$

It is to be noted that while the joint density of $z_1,\ldots,z_n$ given $\theta_{i_1},\ldots,\theta_{i_n}$ factorizes to $\Pi_{j=1}^{n} f(z_j|\theta_{i_j},\psi_{i_j},\pi_{i_j})$, this will not be generally true for the predictive density; i.e.,

$$f(z_1,\ldots,z_n|X,\psi,\pi_{i_1},\ldots,\pi_{i_n}) \neq \prod_{j=1}^{n} f(z_j|X,\psi_{i_j},\pi_{i_j}). \tag{2.21}$$

Hence the results of a joint allocation will be in principle different from the previous type, which we may refer to as a marginal allocation, although perhaps not too often in practice.

It is sometimes convenient to write

$$\Pr\{z_1 \in \pi_{i_1},\ldots,z_n \in \pi_{i_n}|X,\psi,q\} = \Pr\{Z_1 \in \pi_1,\ldots,Z_k \in \pi_k|X,\psi,q\}, \tag{2.22}$$

where Z_i represents the set of n_i observations assumed from π_i and $\Sigma_{i=1}^{k} n_1 = n$, since the set of observations $z_1,\ldots,z_n$ is apportioned among the k populations such that n_i belong to π_i. The reason for using (2.22) is that under certain conditions we do have a useful factorization such that

$$\Pr\{Z_1 \in \pi_1,\ldots,Z_k \in \pi_k|X,\psi,q\} = \prod_{j=1}^{k} \Pr\{Z_j \in \pi_j|X,\psi,q\}. \tag{2.23}$$

Another form of predictive classification would be one wherein diagnoses or allocations need be made as soon as possible, i.e., as soon as z_i is observed. Hence, if $z_1, z_2, \ldots$ are observed sequentially, we may wish, when we are ready to observe and classify z_n, to make our allocation as precise as possible by incorporating the previous observations $z_1, \ldots, z_{n-1}$ into our predictive apparatus. We need now compute the sequential predictive density of z_n on the hypothesis that it belongs to π_i conditional on ψ and on the observations X (whose population origin is known), and on the observations $z_1, \ldots, z_{n-1}$ (whose population origin is uncertain). We then obtain the sequential predictive density of z_n on the hypothesis that it belongs to π_i.

$$f(z_n|X,\psi,z_1,\ldots,z_{n-1},\pi_i)$$

$$\propto \sum_{i_{n-1}=1}^{k} \cdots \sum_{i_1=1}^{k} q_{i_1} \cdots q_{i_{n-1}} f(z_1,..,z_n|X,\psi,\pi_{i_1},..,\pi_{i_{n-1}},\pi_i), \qquad (2.24)$$

i.e., a mixture of joint predictive densities with z_n assumed from π_i. Further,

$$\Pr\{z_n \in \pi_i|X,\psi,z_1,\ldots,z_{n-1}\} \propto q_i f(z_n|X,\psi,z_1,\ldots,z_{n-1},\pi_i). \qquad (2.25)$$

This same result can also be obtained from the product of the likelihoods and the prior density,

$$L(X|\theta,\psi)L(z_1,\ldots,z_{n-1}|\theta,\psi)g(\theta|\psi) \propto p(\theta|X,\psi,z_1,\ldots,z_{n-1}), \qquad (2.26)$$

where

$$L(z_1,\ldots,z_{n-1}|\theta,\psi) = \prod_{j=1}^{n-1} \sum_{i_j=1}^{k} q_{i_j} f(z_j|\theta_{i_j},\psi_{i_j})$$

and finally,

$$f(z_n|X,\psi,z_1,\ldots,z_{n-1},\pi_i) = \int f(z_n|\theta_i,\psi_i)p(\theta|X,\psi,z_1,\ldots,z_{n-1})d\theta, \quad (2.27)$$

which is equivalent to (2.24).

3. Multivariate Normal Allocation.

We now illustrate the previous work by applying it to multivariate normal distributions. The usual situation is to assume equal covariance matrices but differing means for the k populations $\pi_1,\ldots,\pi_k$. Hence π_i is represented by a $N(\mu_i,\Sigma)$ distribution with an available training sample $x_{i1},\ldots,x_{iN_i}$, $i = 1,\ldots,k$. We define

$$\bar{x}_i = N_i^{-1}\sum_{j=1}^{N_i} x_{ij} \; ; \; (N_i-1)S_i = \sum_j (x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'$$

$$(N-k)S = \sum_i (N_i-1)S_i, \quad N = \sum_{i=1}^{k} N_i .$$

Using a convenient reference prior for $\mu_1,\ldots,\mu_k$ and Σ^{-1}

$$g(\Sigma^{-1},\mu_1,\ldots,\mu_k) \propto |\Sigma|^{\frac{1}{2}(p+1)} \quad (3.1)$$

we easily obtain, including only relevant constants,

$$f(z|\bar{x}_i, S,\pi_i) \propto \left(\frac{N_i}{N_i+1}\right)^{p/2}\left[1+\frac{N_i(\bar{x}_i-z)'S^{-1}(\bar{x}_i-z)}{(N_i+1)(N-k)}\right]^{-(N-k+1)/2} \quad (3.2)$$

the predictive density of the observation to be allocated. This then is inserted into either (2.6), (2.16) or (2.17) depending on the circumstances involving q and is appropriate for allocating a single new vector observation z_1.

We now assume that we need jointly allocate n new vector observations $z_1,\ldots,z_n$. Letting, as in (2.15), Z_i represent the set of n_i

observations assumed from π_i, $n = \sum_{i=1}^{k} n_i$ we obtain

$$f(z_1,\ldots,z_k|X,\pi_1,\ldots,\pi_k) \propto \left(\prod_{i=1}^{k} \frac{N_i}{N_i+n_i}\right)^{p/2}$$

$$\cdot|(N-k)S+\sum_{i=1}^{k}(z_i-\bar{x}_i e_i')\,\Omega_i(z_i-\bar{x}_i e_i')'|^{-\frac{N+n-k}{2}} \tag{3.3}$$

where $\Omega_i = I + N_i^{-1}e_i e_i'$ and $e_i' = (1,\ldots,1)$ of dimension n_i.
Hence

$$\Pr[z_1 \in \pi_1,\ldots,z_k \in \pi_k|X,q] \propto \left(\prod_{i=1}^{k} q_i^{n_i}\right) f(z_1,\ldots z_k|X,\pi_1,\ldots,\pi_k) \tag{3.4}$$

where again if the q_i so are unknown appropriate substitutes can be found in (2.16), or what follows it.

The observations may in many instances be sequentially obtained and for compelling reasons allocations (diagnoses) made as soon as possible.

Let $z^{(n-1)} = (z_1,\ldots,z_{n-1})$ and Σ' stand for the sum over all assignments of $z_1,\ldots,z_{n-1}$ to $Z_1,\ldots,Z_k$ with z_n always assigned to Z_i and then summed over all partitions of n such that $\sum_{j=1}^{k} n_j$, $n_j \geq 0$, $j \neq i$ and $n_i \geq 1$. Then

$$\Pr[z_n \in \pi_i|X,z^{(n-1)},q] \propto \sum{}' \left(\prod_{j=1}^{k} q_j^{n_j}\right)\left(\prod_{j=1}^{k} \frac{N_j}{N_j+n_j}\right)^{p/2}$$

$$\cdot|(N-k)S + \sum_{i=1}^{k}(z_i-\bar{x}_i e_i')\Omega_i(z_i-\bar{x}_i e_i')'|^{-\frac{N+n-k}{2}} \tag{3.5}$$

for $n = 2,3,\ldots$.

A second case that is also easily managed is the unequal covariance matrix situation. Here π_i is represented by a $N(\mu_i,\Sigma_i)$ distribution $i = 1,\ldots,k$.

Using the same training sample notation as previously and a similar convenient unobtrusive reference prior

$$g(\mu_1,\ldots,\mu_k,\Sigma_1^{-1},\ldots\Sigma_k^{-1}) \propto \prod_{i=1}^{k} |\Sigma_i|^{\frac{1}{2}(p+1)} \tag{3.6}$$

we obtain

$$f(z|\bar{x}_i,s_i\pi_i) \propto \left(\frac{N_i}{N_i+1}\right)^{\frac{p}{2}} \frac{\Gamma\left(\frac{N_i}{2}\right)\left[1+\frac{N_i(\bar{x}_i-z)'S_i^{-1}(\bar{x}_i-z)}{N_i^2-1}\right]^{-N_i/2}}{\Gamma\left(\frac{N_i-p}{2}\right)|(N_i-1)S_i|^{\frac{1}{2}}} \tag{3.7}$$

the predictive density of the observation to be allocated. This is then inserted into the appropriate formula as previously to calculate the posterior probability of z belonging to π_i.

For the joint classification of $z_1,\ldots,z_n$ we obtain as in (2.15) by assigning $z_1,\ldots,z_n$ to $z_1,\ldots,z_k$

$$\Pr[z_1 \in \pi_1,\ldots,z_k \in \pi_k|X,q] \propto \prod_{i=1}^{k} q_i^{n_i} d(z_i|\bar{x}_i e_i',\Omega_i,s_i,N_i-1,n_i,p) \tag{3.8}$$

where $d(\cdot|\cdot)$ represents the determinantal density Geisser (1966),

$$d(Y|\Lambda,\Omega,A,M,m,p) = \frac{(2\pi)^{-pm/2}\, K(p,M)\, |MA|^{M/2}\, |\Omega|^{p/2}}{K(p,m+m)\, |MA+(y-\Lambda)\Omega(y-\Lambda)'|^{(M+m)/2}} \tag{3.9}$$

for $M \geq p$, $m \geq 1$, A is $p\times p$ and positive definite, Ω is $m\times m$ and positive definite, Y and Λ are $p\times m$, and in addition (3.9) is defined as 1 for $m = 0$.

For sequential allocation we obtain, for $n = 2,3,\ldots$

$$\Pr\{z_n \in \pi_i|X,z^{(n-1)},q\} \propto \Sigma' \prod_{j=1}^{k} q_j^{n_j} d(z_j|\bar{x}_j e_j',\Omega_j,s_j,N_j-1,n_j,p) \tag{3.10}$$

A third case of interest, especially in genetic studies of monozygotic addizygotic twins is where π_i is represented by a $N(0,\Sigma_i)$ distribution $i = 1, 2$.

Again assuming a prior of the form

$$g(\Sigma_1^{-1}, \Sigma_2^{-1}) \propto |\Sigma_1\Sigma_2|^{\frac{1}{2}(p+1)} \tag{3.11}$$

we obtain the predictive density of the vector difference of a twin pair,

$$f(z|X,\pi_i) \propto \frac{\Gamma\left(\frac{N_i+1}{2}\right)}{\Gamma\left(\frac{N_i+1-p}{2}\right)} \frac{|N_iT_i|^{N_i/2}}{|N_iT_i+zz'|^{(N_i+1)/2}} \tag{3.12}$$

where $N_iT_i = \sum_{j=1}^{N_i} x_{ij}x_{ij}'$, and x_{ij} represents the vectorial difference between a twin pair. Insertion of (3.12) into the appropriate formula yields the posterior probability of the new twin pair being either monozygotic or dizygotic.

For joint classification,

$$Pr\{z_1 \in \pi_1, z_2 \in \pi_2|X,q\} \propto \prod_{i=1}^{2} q_i^{n_i} d(z_i|0,I,T_i,N_i,n_i,p)$$

and for sequential allocation

$$Pr\{z_n \in \pi_i|X,z^{(n-1)},q\} \propto \sum' \prod_{j=1}^{2} q_j^{n_j} d(z_j|0,I,T_j,N_j,n_j,p).$$

The material in this and the previous section is derived from Geisser (1964, 1965, 1966) and Geisser and Cornfield (1963), Geisser and Desu (1968, 1973).

4. Bayesian Separation

A second goal in discrimination studies is to identify and utilize in some parsimonious manner the manifest features that separate the various populations. Here there are no new observations that require allocation. The stress is on throwing some light on scientific, technical or social issues.

One defines a class $\mathcal{D}(z)$ of discriminants, and some measure of spread amongst the populations and then selects some minimal set of discriminants that maximizes the spread given the constraints. The technique appears to work best when the p-dimensional multivariate populations can be assumed to have approximately the same covariance matrix Σ and differing mean vectors $\mu_1,\ldots,\mu_2$, and exhibit roughly the kind of symmetry possessed by multivariate normal densities. Hence the major source of their differences is their location. Fisher (1936) found the set of linear combinations $c'z$ which maximized pairwise the distance function

$$\frac{c'(\mu_i-\mu_j)(\mu_i-\mu_j)c}{c'\,\Sigma\, c} \qquad i,j = 1,\ldots,k. \tag{4.1}$$

Let

$$\Lambda\Lambda' = \sum_{i=1}^{k} (\mu_i-\bar{\mu})(\mu_i-\bar{\mu})' \tag{4.2}$$

be of rank $k - v \leq p$ and Λ is $p \times k - v$ and in particular $v = 1$ if $\mu_1,\ldots,\mu_k$ are linearly independent. The solution then is the set of $k - v$ linear discriminants given by

$$z'\Sigma^{-1}\,\Lambda\, P \tag{4.3}$$

where P is $k-v \times k-v$ orthogonal matrix which reduces $\Lambda'\Sigma^{-1}\Lambda$ to

diag$(\delta_1,\ldots,\delta_{k-v})$ matrix and δ_j are the non-zero roots in descending order of $A'\Sigma^{-1}A$. Fisher's derivations essentially employed Lagrange multipliers. An alternate geometric derivation is given by Dempster (1969). Wilks (1962) obtained these results by maximizing a single measure of spread. A somewhat more general approach using algebraic methods is given by Geisser (1977), who demonstrates that any scalar measure of the spread of the k populations that is increasing in the non-zero roots of $AA'\Sigma^{-1}$ is maximized in an $r \leq p$ dimensional space by that set of $r \leq k-v \leq p$ linear discriminants

$$z'\Sigma^{-1} A P_{(r)} \tag{4.4}$$

where $P'_{(r)} = (P_1,\ldots,P_r)$ are the r column vectors associated with the r largest non-zero roots δ_j of $AA'\Sigma^{-1}$.

The focus here is on the estimation of c. In particular if we are dealing with two populations (3.4) is equivalent to

$$z'\Sigma^{-1}(\mu_1-\mu_2) \tag{4.5}$$

To estimate this quantity in Bayesian manner would generally require a joint posterior distribution for Σ^{-1}, μ_1, and μ_2 and hence precise distributional specifications on π_1 and π_2. However if we take the posterior mean of $z'\Sigma^{-1}(\mu_1-\mu_2)$ as its estimator and assume that $E(\mu_1-\mu_2|\Sigma^{-1})$ is $\bar{x}_1-\bar{x}_2$ and the marginal expectation of Σ^{-1} is S^{-1} where, in terms of the sample values in section 2, $(n_1+n_2-2)S = (n_1-1)S_1 + (u_2-1)S_2$ then we have the result that the Bayesian estimator of $z'\Sigma^{-1}(\mu_1-\mu_2)$ is

$$z'S^{-1}(\bar{x}_1-\bar{x}_2) \ . \tag{4.6}$$

If we make the multivariate normal assumptions of section 2 and also

use the same prior density for Σ^{-1}, μ_1, μ_2 we obtain the result of (4.6). Hence one may obtain for k populations that the estimator for $z'\Sigma^{-1}(\mu_i-\mu_j)$ is $z'S^{-1}(\bar{x}_i-\bar{x}_j)$ and generate the estimator

$$z'S^{-1}\hat{\Lambda}\hat{P} \tag{4.7}$$

of the set of linear discriminants, where $\hat{\Lambda}$ and $\hat{P}$ are obtained from the solution

$$\hat{P}\hat{\Lambda}'S^{-1}\hat{\Lambda}\hat{P} = \text{Diag.} \tag{4.8}$$

5. Allocatory-Separatory Compromises

By an allocatory-separatory compromise we mean that we shall derive the discriminant from allocatory/separatory considerations and apply it in semi-Bayesian manner for separatory/allocatory purposes.

For the sake of simplicity we shall confine ourselves to the two population case π_1 or π_2 as there is no intrinsic difficulty in extending it to the case of k populations. Assume now that π_i is specified by density $f(\cdot|\theta_i,\pi_i)$ suppressing the known parameter ψ_i. For purposes of allocation we obtain

$$\rho = \frac{f(z|\theta_i,\pi_1)}{f(z|\theta_i,\pi_2)} \begin{matrix} \geq \\ < \end{matrix} q_2q_1^{-1} \begin{matrix} \text{allocate } z \text{ to } \pi_1 \\ \text{allocate } z \text{ to } \pi_2 \end{matrix} \tag{5.1}$$

or

$$\begin{matrix} h(\rho) \geq h(q_2q_1^{-1}) & \text{allocates } z \text{ to } \pi_1 \\ < h(q_2q_1^{-1}) & \text{allocates } z \text{ to } \pi_2 \end{matrix} \tag{5.2}$$

where $h(\rho)$, any monotone function, equally serves as an identical allocator. We could also consider ρ or $h(\rho)$ as a separatory function derived initially from allocatory considerations. In frequentist theory

ρ depends on θ so an estimate of ρ is obtained by plugging an estimate for the set of parameters θ obtained from the training sample employing some "optimal" estimation property. However as is usually the case these optimal properties will not ordinarily be invariant under monotone transformations, e.g. mean squared error. A way around this dilemma which preserves the invariance of the allocation rule is to use an estimator $\hat{\theta}$ such that $\hat{h}(\rho) = h(\hat{\rho})$, in particular the maximal likelihood estimator of ρ. Of course, for purposes of allocating one might attempt to derive an estimator of h (or better a rule) which minimized future errors of allocation. However this in general cannot be achieved for all θ when θ must be estimated from a training sample. A semi-Bayesian approach to estimating ρ or $h(\rho)$ is fraught with some of the same difficulties. For example minimizing posterior squared error implies that the Bayesian estimator is $E_\theta(h(\rho))$, where the expectation is taken over the posterior distribution of θ. However this estimator will not in general be equal to $h(E_\theta(\rho))$, thus this loss function does not possess the invariance property. In order to retain the invariance property, one could use the posterior median of $h(\rho)$. In practice this turns out to be a rather difficult computation. Hence one settles for a convenient and simple function $h(\rho)$ and calculates its posterior expectation, Geisser (1967), Enis and Geisser (1970). Note that we started from an allocatory point of view and obtained a separatory function. One sometimes also is interested in finding the allocatory properties of such a separatory function or more generally a Bayesian analysis of error rates of any proposed separatory discriminant.

Another semi-Bayesian way of proceeding is to start from the predictive density functions which are the prime ingredients of allocatory rules

and then define a class of separatory discriminants selecting that one which minimizes the total error of classification with respect to the predictive distributions, Enis and Geisser (1974). This then would be an all purpose discriminant having both good separatory and allocatory properties. This approach modifies the optimal Bayesian allocatory discriminant, which is

$$r(z) = \frac{q_1 f(z|X,\pi_1)}{q_2 f(z|X,\pi_2)} \quad , \tag{5.3}$$

by introducing a constraint on the form of the discriminant. Define $W(z, c)$ as a member of the class $\mathcal{D}(z)$ where c is a set unknown constants such that

$$\begin{aligned} W(z,c) \geq 0 \quad & \text{allots } z \text{ to } \pi_1 \\ < 0 \quad & \text{allots } z \text{ to } \pi_2 \end{aligned} \tag{5.4}$$

Then minimize

$$\varepsilon(c) = q_1\varepsilon_1(c) + q_2\varepsilon_2(c) \tag{5.5}$$

with respect to c where

$$\begin{aligned} \varepsilon_1(c) &= \Pr[W < 0 | X,\pi_1] \\ &= \int_{-\infty}^{0} f(w|X,\pi_1)dw \\ \varepsilon_2(c) &= \Pr[W \geq 0|X,\pi_1] \\ &= \int_{0}^{\infty} f(w|X,\pi_2] \end{aligned} \tag{5.6}$$

where $f(w|X,\pi_i)$ is the predictive density of W derived from the predictive density of Z given π_i.

This provides us with a Bayesian discriminant of a stipulated form that is optimal with respect to error rates. This compromises the form of the discriminant with an allocatory requirement.

6. Semi-Bayesian Multivariate Normal Applications

In the multivariate normal case with equal covariance matrices interest has generally focussed on the linear discriminant

$$U = \log \rho = [z - \tfrac{1}{2}(\mu_1+\mu_2)]'\Sigma^{-1}(\mu_1-\mu_2) \tag{6.1}$$

with the accompanying allocatory rule

$$\begin{aligned} U \geq \log r \quad &\text{assigns } z \text{ to } \pi_1 \\ U < \log r \quad &\text{assigns } z \text{ to } \pi_2 \end{aligned} \tag{6.2}$$

The usual frequentist estimator of u is the sample linear discriminant

$$V = [z - \tfrac{1}{2}(\bar{x}_1+\bar{x}_2)]'S^{-1}(\bar{x}_1-\bar{x}_2) \tag{6.3}$$

obtained by substituting the usual estimators for μ_1, μ_2, and Σ in U. The actual allocation rule derived from the training sample then is as follows:

$$\begin{aligned} V \geq \log r \quad &\text{assigns } z \text{ to } \pi_1 \\ V < \log r \quad &\text{assigns } z \text{ to } \pi_2 . \end{aligned} \tag{6.4}$$

The first thing we note is that we can provide a Bayesian estimator of U by calculating its posterior expection with regard to μ_1, μ_2 and Σ for fixed z. For the particular prior distribution used previously

$$E(U|z) = V + \tfrac{1}{2}p(N_2^{-1}-N_1^{-1}) \tag{6.5}$$

is a Bayesian estimator of U and in terms of its use as a separatory discriminant is virtually identical to V because for separatory purposes the constant displacement is more or less irrelevant. Frequentists also use V in its allocatory mode, as determined by (6.4). There it

appears that the Semi-Bayes approach may yield a rather slight improvement that diminishes with increasing sample sizes and decreasing difference between sample sizes, in terms of frequentist error rates.

Although the frequentist theory of allocation concerns itself with a number of different error rates and their estimators, Hills (1966), we shall only discuss the two most important ones. Now the optimal errors of classification are given as

$$\Pr[U < \log r \,|\, \mu_1, \mu_2, \Sigma^{-1}, \pi_1] = \varepsilon_1 = \int_{-\infty}^{\tau_1} \phi(v)dv = \Phi(\tau_1), \tag{6.6}$$

$$\Pr[U > \log r \,|\, \mu_1, \mu_2, \Sigma^{-1}, \pi_2] = \varepsilon_2 = \int_{\tau_2}^{\infty} \phi(v)dv = 1-\Phi(\tau_2) \tag{6.7}$$

where $\phi(v) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}v^2}$ is the standard normal density and

$$\tau_1 = (\log r - \tfrac{1}{2}\alpha)/\alpha^{\frac{1}{2}}, \tau_2 = (\log r + \tfrac{1}{2}\alpha)/\alpha^{\frac{1}{2}}, \tag{6.8}$$

$$\alpha = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) . \tag{6.9}$$

A frequentist estimator of the optimal errors employs V for U and for α substitutes

$$\hat{\alpha} = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) = Q \tag{6.10}$$

A Bayesian estimator for ε_i can, in principle, be obtained by calculating $E(\varepsilon_i)$. This is a rather difficult calculation and an approximation is available. Let $c = N_1N_2/(N_1+N_2)$ and $\nu = N_1+N_2-2$ then

$$E(\varepsilon_1) \cong \Phi\left(\frac{\log r - \tfrac{1}{2}(pc^{-1} + Q)}{[pc^{-1} + (1+pc^{-1})Q]^{\frac{1}{2}}}\right) \tag{6.11}$$

and with increasing sample size is

$$E(\varepsilon_1) \cong \Phi\left(\frac{\log r - \tfrac{1}{2}Q}{Q^{\frac{1}{2}}}\right) = \hat{\varepsilon}_1 \tag{6.12}$$

Further one can obtain,

$$P(\varepsilon_1) = Pr[\varepsilon_1 \leq b] \cong 1 - F_\alpha\left[\frac{4c(p+cQ)(\Phi^{-1}(b))^2}{p+cQ+\nu^{-1}(cQ)^2}\right] \tag{6.13}$$

where $\nu = N_1+N_2-2$ and $F_d(\cdot)$ is the distribution function of a chi-squared random variable with $d = (p+cQ)^2/(p+cQ+(cQ)^2\nu^{-1})$ degrees of freedom. A similar result is available for ε_2.

The estimation of $\varepsilon = q_1\varepsilon_1 + q_2\varepsilon_1$, the total optimal error rate is useful as a guide to the optimal discriminatory power of the variables used for allocation. If the estimate of ε indicates that ε is larger than the accuracy required of the allocation procedure one would search for additional or another set of variables that would diminish the total error rate. If (6.5) is used for allocation in place of (6.4) then one replaces $\log r$ by $\log r - \frac{1}{2}p(N_2^{-1}-N_1^{-1})$ in the Bayesian estimators of (6.11) and (6.12).

From the practical point of view the error rates that are most important are those that are actually incurred when using the sample discriminant V on future observations. These actual errors are defined as

$$Pr(V < \log r|\mu_1,\mu_2,\Sigma,\pi_1) = \beta_1 = \int_{-\infty}^{\theta_1}\phi(v)dv = \Phi(\theta_1), \tag{6.14}$$

$$Pr(V > \log r|\mu_1,\mu_2,\Sigma,\pi_2) = \beta_2 = \int_{\theta_2}^{\infty}\phi(v)dv = 1 - \Phi(\theta_2), \tag{6.15}$$

for the fixed values $\bar{x}_1$, $\bar{x}_2$ and S where

$$\theta_1 = \{[\tfrac{1}{2}(\bar{x}_1+\bar{x}_2) - \mu_1]'S^{-1}(\bar{x}_1-\bar{x}_2) + \log r\} \cdot[(\bar{x}_1-\bar{x}_2)'S^{-1}\Sigma S^{-1}(\bar{x}_1-\bar{x}_2)]^{-\frac{1}{2}}, \tag{6.16}$$

$$\theta_2 = \{[\tfrac{1}{2}(\bar{x}_1+\bar{x}_2) - \mu_2]'S^{-1}(\bar{x}_1-\bar{x}_2) + \log r\} \cdot[(\bar{x}_1-\bar{x}_2)'S^{-1}\Sigma S^{-1}(\bar{x}_1-\bar{x}_2)]^{-\frac{1}{2}} \tag{6.17}$$

and θ_1 and θ_2 are random variables that are functions of μ_1, μ_2 and Σ. Hence we have defined β_1 and β_2 as functions of the random variables μ_1, μ_2, Σ for fixed values of $\bar{x}_1$, $\bar{x}_2$ and S which differs from the sampling interpretation where β_1 and β_2 are considered either as functions of the fixed parameters μ_1, μ_2, and Σ obtained from the unconditional sampling distribution of V in terms of the random variables $\bar{x}_1$, $\bar{x}_2$, and S, or defined as functions of the random variables $\bar{x}_1$, $\bar{x}_2$, and S. Although the exact posterior distribution of β_1 both jointly or marginally Geisser (1967) can easily be found, a convenient and rather good approximation is obtained as

$$\Pr[\beta_1 \leq b] \cong \Phi\left(\frac{\Phi^{-1}(b) - A_1}{(N_1^{-1} + B_1)^{\frac{1}{2}}}\right) \tag{6.18}$$

where

$$A_1 = \left(\frac{\nu - p + \frac{1}{2}}{Q\nu}\right)^{\frac{1}{2}} (\log r - \tfrac{1}{2}Q)$$

$$B_1 = [\log r - \tfrac{1}{2}Q]^2/2\nu Q$$

$$\Pr[\beta_2 \leq b] = 1 - \Phi\left(\frac{\Phi^{-1}(1-b) - A_2}{(N_2^{-1} + B_2)^{\frac{1}{2}}}\right) \tag{6.19}$$

$$A_2 = \left(\frac{\nu - p + \frac{1}{2}}{Q\nu}\right)^{\frac{1}{2}} (\log r + \tfrac{1}{2}Q)$$

$$B_2 = (\log r + \tfrac{1}{2}Q)^2/2\nu Q \ .$$

A Bayesian estimator of β_i, $E(\beta_i)$, is also the unconditional predictive probability

$$E(\beta_1) = \Pr[V \leq \log r | X, \pi_1] \tag{6.20}$$

$$E(\beta_2) = \Pr[V > \log r | X, \pi_2]. \tag{6.21}$$

The argument runs as follows:

$$E(\beta_1) = \Pr[V \leq \log r | \mu_1, \mu_2, \Sigma, \pi_1] p(\mu_1, \mu_2, \Sigma^{-1} | X) d\mu_1 d\mu_2 d\Sigma^{-1}$$

$$= \int_{-\infty}^{\log r} f(V | \mu_1, \mu_2, \Sigma, \pi_1) p(\mu_1, \mu_2, \Sigma^{-1} | X) d\mu_1 d\mu_2 d\Sigma^{-1} dV \tag{6.22}$$

where $f(V|\mu_1,\mu_2,\Sigma,\pi_1)$ represents the conditional density of V. Hence

$$E(\beta_1) = \int_{-\infty}^{\log r} f(V|X,\pi_1)dV = \Pr[V \leq \log r|X,\pi_1] \tag{6.23}$$

where $f(V|X,\pi_1)$ represents the unconditional or predictive density of V.

Thus we can obtain

$$E(\beta_1) = \Pr[t_{\nu+1-p} \leq (\log r - \tfrac{1}{2}Q)[\nu(N_1+1)Q/(\nu+1-p)N_1]^{-\frac{1}{2}}] \tag{6.24}$$

which may be evaluated directly from tables of the t-distribution. Similarly

$$E(\beta_2) = \Pr[V > \log r | X, \pi_2]$$

$$= \Pr[t_{\nu+1-p} > (\log r + \tfrac{1}{2}Q)[\nu(N_2+1)Q/\nu+1-p)N_2]^{-\frac{1}{2}}]. \tag{6.25}$$

In practice if an investigator is satisfied with the estimate of the optimal error ϵ, then he can compute his estimates of β_1 and β_2. If they are larger then he can tolerate he should collect larger sample sizes since $\beta \to \epsilon$ from above as the sample sizes increase. Of course all of this is prior to obtaining the observations to be allocated since once they are in hand the only relevant calculation for the Bayesian is the posterior

probability that $z \in \pi_i$ or the allocatory decision for that observation. The optimal and actual probability of correct allocation $1 - \varepsilon$, and $1 - \beta$ refer only to the long run frequency of future allocations using the discriminant from a hypothetically infinite sample in the first case and the actual sample in hand in the second case. A more detailed exposition with other results can be found in Geisser (1967, 1970).

Another semi-Bayesian approach would be to find the linear discriminant $W(z) = a'z - b$ such that if

$$\begin{aligned} W(z) \geq 0, &\quad \text{assign } z \text{ to } \pi_1 \\ W(z) < 0, &\quad \text{assign } z \text{ to } \pi_2, \end{aligned} \tag{6.26}$$

where $a' = [a_1, \ldots, a_p]$ is a nonnull vector, and b an arbitrary scalar, such that for variations in a and b the total predictive probability of correct allocation is maximized. The solution obtained by Enis and Geisser (1974) is termed the optimal predictive linear discriminant,

$$W_0(z) = a_0'z - b_0 \tag{6.27}$$

where

$$a_0 = S^{-1}(\bar{x}_1 - \bar{x}_2) \tag{6.28}$$

$$b_0 = (RK_1^2 - K_2^2)^{-1}\{(\bar{x}_1 - \bar{x}_2)'S^{-1}(RK_1^2\bar{x}_1 - K_2^2\bar{x}_2) - Q^{\frac{1}{2}}\omega^{\frac{1}{2}}\}, \tag{6.29}$$

$$\omega = QRK_1^2K_2^2 - \nu(R - 1)(RK_1^2 - K_2^2),$$

$$R = \left(\frac{q_2K_2}{q_1K_1}\right)^{2/(\nu+1)},$$

and

$$K_i = \left[\frac{\nu N_i}{(N_i+1)(\nu+p-1)}\right]^{\frac{1}{2}}.$$

First we note that for purely separatory purposes the constant is irrelevant and again we obtain Fisher's linear discriminant function.

For allocatory purposes the constant b_0 is relevant and may yield a rather slight error rate improvement over V or $V + p(N_2^{-1}-N_1^{-1})$. But note that W will be globably optimal iff $RK_1^2 = K_2^2$ since this is equivalent to r(z), the optimal posterior discriminant, which under these circumstances an appropriate h(r) becomes linear as well. If $q_1 = q_2$ and $N_1 = N_2$ than all methods thus far essentially yield V.

When $\Sigma_1 \neq \Sigma_2$

the optimal discriminant is the quadratic

$$U = \tfrac{1}{2}\{\log|\Sigma_1^{-1}\Sigma_2| + (z-\mu_2)'\Sigma_2^{-1}(z-\mu_2) - (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1)\} \quad . \quad (6.30)$$

Error rates become much more difficult to compute under these circumstances. It is however interesting to note that the usual estimator of U, namely

$$V = \tfrac{1}{2}\{\log|S_1^{-1}S_2| + (z-\bar{x}_2)'S_2^{-1}(z-\bar{x}_2) - (z-\bar{x}_1)'S_1^{-1}(z-\bar{x}_1)\} \quad (6.31)$$

is also very nearly achieved as the posterior expectation of U. Enis and Geisser (1970) show that

$$E(U|z) = V + h(p, N_1, N_2) \quad (6.32)$$

where

$$h(p,N_1,N_2) = \tfrac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{p}(-1)^i\{\log(N_i-1)+N_i^{-1}-\Psi[\tfrac{1}{2}(N_i-j)]\} \quad . \quad (6.33)$$

and $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the psi(digamma) function. Note as N_1 and N_2 increase $h \to 0$ and in particular when $N_1 = N_2$, $h \equiv 0$. Thus V differs from the posterior expectations by at most a negligible quantity.

The optimal predictive discriminant is

$$r(z) = \frac{f(z|\bar{x}_1, S_1, \pi_1)}{f(z|\bar{x}_2, S_2, \pi_2)} \gtrless q_2 q_1^{-1} \tag{6.34}$$

as defined in (3.7) and is a rather complicated function of z and no $h(r)$ emerges that will simplify it. One could attempt to derive the optimal predictive quadratic discriminant but in general this is quite difficult to obtain.

7. Semi-Bayesian Sample Reuse Selection and Allocation

In many problems we cannot always formulate definitively the density function for π_1 and π_2. For example in certain situations we may be uncertain as to whether we are dealing with two normal populations with differing means and either the same or differing covariance matrices. Hence often to the problem of allocation there is added an uncertainty regarding the specification. More generally, suppose that $f(\ |\theta_i, \pi_i, \omega)$ the basic density is now indexed by the double designator $\omega \in \Omega$ which jointly specifies a pair of densities for π_1 and π_2 and is assumed subject to a probability function $g(\omega)$. A complete Bayesian solution for the allocation of z (Geisser, 1979) maximizes

$$\max_i \Pr[z \in \pi_i | X, q_i] \propto q_i E_\omega f(z|X,w,\pi_i) f(X_1,X_2|\pi_1,\pi_2,\omega) \tag{7.1}$$

where the expectation is over $g(\omega)$, X_i represents the set of observations from π_i,

$$f(z|X,\omega,\pi_i) = \int f(z|\theta_i,\omega,\pi_i) g(\theta|X,\omega) d\theta \tag{7.2}$$

where f, the sampling density of z and g, the posterior density of θ in the integrand are now indexed by ω which specifies the assumed population and

$$f(X_1,X_2|\omega,\pi_1,\pi_2) = \int g(\theta|\omega) \prod_{i=1}^{2} \prod_{j=1}^{N_i} f(x_{ij}|\theta_i,\omega,\pi_i) d\theta \tag{7.3}$$

This full Bayesian approach requires a body of prior knowledge that is often unavailable and may be highly sensitive to some of these assumptions.

We shall present here only one of a series of data analytic techniques given by Geisser (1979) which selects a single $\omega = \omega^*$ to be used for

allocation rather than the Bayesian averaging. It is a technique which combines Bayesian, frequentist and sample reuse procedures.

Let

$$L(\omega) = \prod_{i=1}^{2} \prod_{j=1}^{N_i} f(x_{ij} | X_{(ij)}, \omega, \pi_i), \tag{7.4}$$

the product of reused predictive densities, where $X_{(ij)}$ is the set of observations X with x_{ij} deleted and f is the same form as (7.2); i.e., x_{ij} replaces z and $X_{(ij)}$ replaces X. Choose ω^* according to

$$\max_{\omega} g(\omega)\, L(\omega),$$

and then use that π_1 and π_2 specified by ω^* in an allocatory or separatory mode.

As an example suppose $\omega = \omega_1$ specified that π_i is $N(\mu_i, \Sigma)$ and $\omega = \omega_2$ specified that π_i is $N(\mu_i, \Sigma_i)$ respectively.

Under ω_1

$$L(\omega_1) = \prod_{i=1}^{2} \prod_{j=1}^{N_i} f(x_{ij} | \bar{x}_{i(j)}, S_{(ij)}, N_i-1, N-1, \omega_1, \pi_i) \tag{7.5}$$

where the density f is given by (3.2) with z, $\bar{x}_i$, S, N_i, and N replaced by x_{ij}, $\bar{x}_{i(j)}$, $S_{(ij)}$, N_i-1 and $N-1$ respectively; $\bar{x}_{i(j)}$ and $S_{(ij)}$ being the sample mean and pooled covariance matrix with x_{ij} deleted.

Under ω_2

$$L(\omega_2) = \prod_{i=1}^{2} \prod_{j=1}^{N_i} f(x_{ij} | \bar{x}_{i(j)}, S_{i(j)}, N_i-1, \omega_2, \pi_i) \tag{7.6}$$

where the density f is given by (3.7) with z, $\bar{x}_i$, S_i and N_i replaced by x_{ij}, $\bar{x}_{(ij)}$, $S_{i(j)}$, and N_i-1 respectively and $S_{i(j)}$ being

the sample covariance matrix calculated from X_i with x_{ij} deleted. The choice of ω^* now rests with

$$\max_i g(w_i) L(w_i), \quad i = 1, 2. \tag{7.7}$$

One then uses the ω^* specification for allocation or separation.

8. Other Areas

Most of the current work in separatory discriminants has been linear mainly because of convenience and ease of interpretation. However it would be desirable to consider other functional discriminants as there are situations where the natural discriminants are quadratic.

There is also another useful model wherein the so-called populations or labels have some underlying continuous distribution, but one can only observe whether π is in a set S_i where $S_1,\ldots,S_k$ exhaust the range of π, see, for example Marshall and Olkin (1968). In the previous case $\pi = \pi_i$ was synonomous with S_i and the distribution involved only the discrete probabilities q_i. However this case involves more structure and requires a more delicate Bayesian analysis. Work in this area is currently in progress.

References

Dempster, A.P. (1969), Elements of Continuous Multivariate Analysis, Addison-Wesley, Reading, Massachusetts.

Desu, M.M. and Geisser, S. (1973), Methods and applications of equal-mean discrimination, Discriminant Analysis and Applications, edited by T. Cacoullos, Academic Press, New York, pp. 139-161.

Enis, P. and Geisser, S. (1970), Sample discriminants which minimize posterior squared error loss, South African Statist. J., 4, pp. 85-93.

Enis, P. and Geisser, S. (1974), Optimal predictive linear discriminants, Ann. Statist., 2, 2, pp. 403-410.

Fisher, R.A. (1936), The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7, pp. 179-188.

Geisser, S. (1964), Posterior odds for multivariate normal classification, J.R Statist. Soc. B, 1, pp. 69-76.

Geisser, S. (1965), Bayesian estimation in multivariate analysis. Ann. Math. Statist. 36, 150-159.

Geisser, S. (1966), Predictive discrimination, Multivariate Analysis, edited by P. Krishnaiah, Academic Press, New York, pp. 149-163.

Geisser, S. (1967), Estimation associated with linear discriminants, Ann. Math. Statist. 38, pp. 807-817.

Geisser, S. (1970), Discriminatory practices, Bayesian Statistics, edited by D. Meyer and R.C. Collier, Peacock, Illinois, pp. 57-70.

Geisser, S. (1977), Discrimination, allocatory and separatory, linear aspects, Classification and Clustering, edited by J. Van Ryzin, Academic Press, New York, pp. 301-330.

Geisser, S. (1979), Sample reuse selection and allocation criteria, Multivariate Analysis V, edited by P. Krishnaiah, North-Holland Publishing Co., Amsterdam, in press.

Geisser, S. and Cornfield, J. (1963), Posterior distributions for multivariate normal parameters, J. Roy. Statist. Soc. Ser. B 25, 368-276.

Geisser, S. and Desu, M.M. (1968), Predictive zero-mean uniform discrimination, Biometrika, 55, 3, pp. 519-524.

Hills, M. (1966), Allocation rules and their error rates, J.R. Statist. Soc. B, 28, pp. 1-31.

Marshall, A.W. and Olkin, I. (1968), A general approach to some screening and classification problems. J.R. Statist. Soc., B, 3, pp. 407-443.

Wilks, S.S. (1962), Mathematical Statistics, John Wiley and Sons, New York.